Attention is Case-Sensitive

Maximilian Dillitzer[1,3,*], Tin Stribor Sohn[2,3,*], Jason Corso[4], and Michael Auerbach[1]

[1] University of Applied Science Esslingen, Esslingen, Germany
[2] Karlsruhe Institute of Technology, Karlsruhe, Germany
[3] Dr. Ing. h.c. F. Porsche AG, Weissach, Germany
[4] University of Michigan, Ann Arbor MI 48109, USA

Abstract. In human visual perception, uppercase lettering serves as a natural salience cue that captures attention within lowercase text. In this paper, we present a systematic empirical characterization study revealing that Large Language Models (LLMs) exhibit an analogous property: letter casing modulates internal attention allocation. Through analysis across 13 models—nine LLMs and four Vision-Language Models (VLMs)—with diverse tokenization schemes, we show that formatting target information in alternating or uppercase against a lowercase context concentrates attention on those textual spans. In text this effect is universal, holding across every evaluated non-reasoning model. We frame it as a previously under-explored latent property of pretrained transformers rather than a prescriptive method. Our investigation reveals a central attention–performance divergence: while this "casing effect" robustly shifts attention, its impact on downstream accuracy is non-trivial—increased concentration does not inherently improve task accuracy and, in high-entropy contexts like alternating case, can degrade it. We further identify a boundary condition: the deliberative "thinking" phase in reasoning models acts as a semantic buffer that mitigates typographic sensitivity in text. Extending the study to VLMs, we find the effect transfers partially: the same prompt-side casing reorganizes cross-modal attention along two coupled axes—predominantly a macroscopic disengagement from the image toward the text prompt, and secondarily a concentration of the residual visual attention on the target region. By isolating casing as a zero-shot mechanism for attention steering that requires no model access or fine-tuning, we provide a new foundational understanding of how pretraining internalizes typographic emphasis.

Keywords: Attention Allocation · Language Models · Token Analysis

1 Introduction

Human visual perception exhibits a well-documented sensitivity to typographic variation: uppercase lettering naturally draws attention within predominantly

[*] Equal contribution.

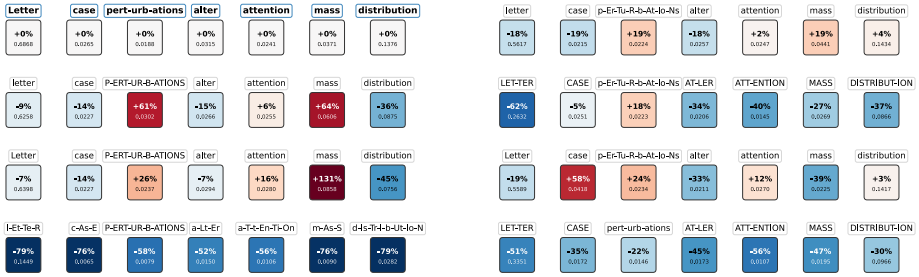


Fig. 1: Letter case as an attentional attractor. LLMs (here, Qwen2.5-7B-Instruct) systematically shift internal attention weights based on typographic casing, even when semantic content remains identical. We show this property is inherent to the model's pretraining and impacts internal state dynamics. Visualizations represent *relative attention mass* compared to a baseline reference sentence (top-left), which is calculated as the *mean attention weight per token* aggregated at the word level.

lowercase text [9,21,48]. Electrophysiological evidence further supports the pre-lexical nature of case-based attentional modulation, establishing case as a surface feature that guides human focus [11,55]. This phenomenon, rooted in the visual salience of capital letters against their lowercase counterparts, has long been exploited in typography, advertising, and interface design to manipulate cognitive priority without altering semantic content [23].

This human perceptual mechanism raises a compelling question for artificial language understanding: do Large Language Models (LLMs), despite operating on abstract token representations rather than visual stimuli, exhibit analogous case-sensitive attention patterns? While case variation is conventionally treated as a passive artifact of tokenization, we hypothesize that LLMs internalize typographic case as a latent importance signal during pretraining. In this paper, we present a systematic empirical study revealing that letter casing is not merely a formatting choice but a fundamental property that modulates internal attention allocation within the transformer architecture.

Our investigation focuses on characterizing this "casing effect" by isolating letter case as the *independent variable* at the pre-tokenization level while strictly preserving word order and semantic context (*cf*. Fig. 1). By analyzing seven non-reasoning and two reasoning LLMs alongside two non-reasoning and two reasoning Vision-Language Models (VLMs) with diverse scales and tokenizer types, we observe that formatting target information in alternating or uppercase against a lowercase context consistently concentrates internal attention weights on those text spans and corresponding image regions.

Importantly, we frame this discovery as an inherent behavioral characteristic of pretrained LLMs and VLMs rather than a prescriptive steering method intended to optimize benchmark performance. Our results show that while the shift in attention is robust and universal, its impact on downstream performance is non-trivial: increased attention concentration can lead to both performance gains

and degradations. This suggests that while case acts as a powerful attentional attractor, "more attention" does not inherently equate to better performance. Notably, we find that the deliberative reasoning process in reasoning models acts as a semantic realignment layer in text that filters out these typographic cues in favor of logical coherence. Crucially, we approach experiments with zero model access—no fine-tuning, adapter insertion, or runtime introspection—operating entirely through prompt engineering (*i.e.*, only changing letter case). Our work makes the following contributions to the foundational understanding of LLM and VLM properties:

- We provide empirical evidence that LLMs and VLMs exhibit a "casing-salience" property, where attention is meaningfully modulated by typographic case, mirroring human visual salience effects across diverse architectures.
- We demonstrate the robustness of this property across seven non-reasoning LLMs, proving the effect persists regardless of tokenization scheme or scale.
- We observe that this pattern also applies to multimodal settings, where image attention in VLMs is steered by letter casing of the textual description.
- We conduct a causal analysis using adversarial de-emphasis (*e.g.*, uppercasing context while lowercasing targets), confirming that the observed attention shifts are a direct response to orthographic variation.
- We reveal a complex relationship between attention allocation and downstream performance, showing that while case can be utilized to influence model focus, the resulting impact on task accuracy is highly context-dependent.
- We identify the "reasoning buffer" as a boundary condition in LLMs, where reasoning mitigates the influence of surface-level typographic attractors.

By characterizing these latent attentional levers, we move beyond viewing case as a tokenization nuance and establish it as a surface-level property that can be utilized for future research in attention steering and model interpretability.

2 Related Work

The internal attention mechanisms of LLMs and VLMs are the primary drivers of their reasoning and generative capabilities. Consequently, a vast body of research has emerged dedicated to "attention steering"—the intentional redirection of a model's focus to improve safety, factual recall, or task performance. Literature typically treats attention as a variable manipulated through external interventions. We categorize these into training-time, inference-time, and activation-level modifications, highlighting the gap our empirical study addresses.

2.1 Engineered Attention Steering

Existing strategies for controlling attention typically require deep architectural or computational interventions:

Training-Time and Structural Modifications. Parameter-level interventions such as ROME [28], MEMAT [50], MEND [32], or REMEDI [19] attempt to "hard-code" attention patterns by editing weights to alter factual recall. Likewise, bias-only adaptation techniques train per-layer steering vectors via reinforcement learning, effectively tuning the model for specific reasoning tasks without full fine-tuning [47]. Other approaches, such as Value-State Gated Attention (VGA) [6], propose architectural redesigns to mitigate pathological attention behaviors. While effective, these studies treat attention as a structural feature that must be modified during training to achieve desired behaviors.

Inference-Time Computational Interventions. Recent works like PASTA [59], AutoPASTA [60], and InstABoost [18] introduce mechanisms to rescale attention scores during the forward pass. These methods "steer" the model by manually amplifying the weights of specific tokens (*e.g.*, instruction tokens). Similarly, SEKA [2] edits key-embeddings before the attention computation occurs and Hsieh *et al.* [20] propose positional calibration to subtract a learned baseline from positional attention bias. While these operate on frozen models, they treat attention as a computational resource that requires white-box hooks and runtime recomputation to be redirected.

Activation Engineering. A third category involves injecting "steering vectors" into the model's residual stream to encourage specific traits [25, 44, 49, 53, 54]. These techniques rely on extracting activation patterns from contrastive datasets and adding them to internal layers during inference [10]. These studies focus on manipulating the *content* of the hidden states to influence attention, rather than exploring the model's natural response to input variations.

2.2 Input-Level Influence and Surface Form

Studies on how input formatting affects LLMs and VLMs has largely focused on semantic or structural perturbations. Researchers have documented positional bias [58], sink token effects [26], and other token-level effects [35, 45], or used XML-like delimiters to provide structural cues for instructions [1]. These works generally operate on the level of *semantic priming*—adding tokens or changing order to influence the model but omitting the effects analysis on vision modalities entirely. In contrast, our work focuses on *orthographic variation* that preserves 100% of the lexical and semantic content alongside unchanged visual inputs in VLMs. While prior work treats casing as a variable to be normalized away during tokenization, our study reveals it as a potent, inherent attentional attractor.

2.3 The Gap: Characterizing Latent Case Sensitivity

The critical gap in current research is the lack of understanding regarding the *inherent* sensitivity of pretrained transformers to typographic surface forms. While the aforementioned steering methods aim to *improve performance* through complex interventions, they overlook a fundamental property of models themselves:

1. **Attention vs. Performance**: Prior works assume that redirecting attention to relevant tokens will monotonically improve accuracy. Our empirical analysis of the "casing effect" provides a more nuanced view, showing that while attention is steered by casing, the resulting task performance depends on the "readability" of the pattern (*e.g.*, uppercase vs. alternating case).
2. **Latent Properties**: Most steering research focuses on *prescriptive methods* (how to force the model to look elsewhere). We provide a *descriptive characterization* of a property that already exists in pretrained weights across nine diverse LLMs and four VLMs.

By documenting how letter casing (*i.e.*, the minimal orthographic perturbation) systematically reshapes attention flow, we bridge cognitive typography and transformer interpretability. This work shifts the perspective from "how can we force attention to move" to "what existing features of text naturally attract it".

3 Research Methodology

Inspired by human visual salience effects where uppercase lettering naturally attracts attention within lowercase text, we conduct a *controlled empirical study* to characterize whether letter case functions as an inherent attentional modulator in pretrained LLMs and VLMs. We deliberately adopt a ground-truth-guided protocol to isolate the pure effect of typographic case from confounding factors, such as span prediction errors. Our objective is the foundational validation of this model property; we treat the observed attention shifts as a behavioral characteristic rather than a prescriptive method. Consequently, while we demonstrate how this property can be utilized for steering, a fully automated pipeline for deployment is outside the scope of this characterization study (*cf*. Sec. 5).

Our protocol is entirely training-free and input-driven. We define static capitalization interventions applied *before tokenization*, measure their causal impact on internal attention distribution (textual and visual attention) and downstream task performance under controlled conditions, and verify the universality of these effects across diverse model families.

3.1 Characterization Scope and Causal Isolation

Given raw input text $X = (x_1, \ldots, x_N)$ as a sequence of characters and either a *ground-truth answer span* or a *bounding box label* $A \subseteq X$ provided by benchmark annotations, we construct a typographically transformed version $\tilde{X}$ such that:

(i) $\tilde{X}$ differs from X *only* in letter case, preserving word order and semantic content of the text exactly;
(ii) $\tilde{X}$ applies a *deterministic rule* formatting A in a target case pattern (*e.g.*, uppercase) while the surrounding context follows a complementary pattern;
(iii) when processed by a frozen LLM or VLM f_θ, $\tilde{X}$ yields measurable shifts in cross-layer attention concentration on A (which can be text or text-corresponding image regions) and changes in task accuracy relative to naturally cased X.

This defines a *pre-tokenization causal isolation protocol* which similarly applies for vision tasks incorporating text captions. The ground-truth dependency is intentional: it eliminates noise from auxiliary models or heuristics, allowing us to attribute any observed changes in model state specifically to the typographic variation.

3.2 Target Span Identification Protocol

To ensure reproducibility, target spans are derived strictly from benchmark-provided labels using a two-stage matching protocol:

- **Exact matching (98.2% of samples)**: We locate A in X via case-insensitive matching. The transformation is applied to the matched character span while preserving non-alphabetic characters (punctuation, digits, whitespace).
- **Fuzzy fallback (1.8% of samples)**: If exact matching fails, we apply Levenshtein-distance alignment (threshold ≤ 2 character edits), ensuring alignment fidelity through a 100-sample subset manual verification.

Crucially, these transformations preserve token boundaries; we do not alter whitespace or lexical structure. For inherently mixed-case entities (*e.g.*, "iPhone"), we apply the transformation to the entire span (*e.g.*, "IPHONE") to maintain consistency in the experimental signal.

3.3 Typographic Intervention Formalism

Let $\mathcal{C} : \Sigma \to \{\text{upper}, \text{lower}, \text{title}, \text{alternating}\}$ denote a case assignment function over character alphabet Σ. Our intervention constructs:

$$\tilde{X} = \mathcal{T}_{\mathcal{C}}(X) = \big(c_1(x_1),\ c_2(x_2),\ \ldots,\ c_N(x_N)\big), \tag{1}$$

where $c_i \in \{\text{upper}, \text{lower}, \text{title}, \text{alternating}\}$ is applied per character, satisfying the semantic preservation constraint:

$$\forall i, \quad \texttt{lower}\big(c_i(x_i)\big) = \texttt{lower}(x_i). \tag{2}$$

This ensures lexical identity remains invariant, even if the underlying tokenizer produces a different sequence of token IDs for $\tilde{X}$ compared to X.

Our study tests the hypothesis that pretrained LLMs and VLMs exhibit an inherent sensitivity to case as an importance signal. Specifically, we investigate the following relationships:

$$\text{Att}_A(f_\theta(\tilde{X})) \neq \text{Att}_A(f_\theta(X)) \quad \text{and in text} \quad \text{Acc}(f_\theta(\tilde{X})) \neq \text{Acc}(f_\theta(X)), \tag{3}$$

where $\text{Att}_A(\cdot)$ denotes the mean attention weight over target A across all layers/heads, and $\text{Acc}(\cdot)$ is the task accuracy of the frozen model f_θ. Crucially, in VLMs, the target A generalizes from discrete textual tokens to spatial image regions, specifically mapping to corresponding patches or pixels. Unlike traditional method-driven papers, we do not assume a monotonic improvement; rather, we characterize the *directionality* and *magnitude* of these shifts to reveal the model's internal response to typographic emphasis.

3.4 Cross-Architecture Robustness Protocol

To evaluate the generality of this property, we apply identical interventions across nine non-reasoning models and four reasoning models spanning five families (LLaMA-3 [17], Gemma-2/3/4 [12,51,52], Mistral [22], Qwen2.5/3 [5,38,40], GPT [34]) and various tokenization schemes. No per-model tuning is performed. By measuring relative changes against each model's naturally cased baseline (as in the benchmarks' protocol), we determine if case sensitivity is a universal emergent property of the transformer architecture or a training regime artifact.

3.5 Principles of the Empirical Study

Our experimental protocol is built on three principles:

1. **Causal Attribution**: Isolation of letter case as the independent variable ensures effects are not due to semantic, structural, or visual changes.
2. **Zero-Inference Observation**: We analyze the behavior of frozen, off-the-shelf models to identify properties already present in the weights.
3. **Phenomenological Analysis**: We report both the successes (performance gains) and the failures (performance degradation) of attention concentration to provide a complete picture of the "casing effect" in text.

By framing our research methodology as a characterization study, we provide the groundwork for future work to utilize these latent attentional levers in more complex, automated steering applications.

4 Experiments

This section details our systematic empirical investigation into the property of case-sensitive attention in LLMs and VLMs. Following the protocol in Section 3, our evaluation follows a two-stage paradigm: we first perform phenomenological discovery on pure text, as text serves as the highest-leverage modality for isolating core transformer mechanics [57], and subsequently transfer these foundational insights to cross-modal attention steering experiments in the vision domain. Accordingly, our experiments are designed to: (i) characterize how diverse typographic interventions modulate internal attention allocation; (ii) quantify the causal relationship between case contrast and attention dynamics through systematic ablations; and (iii) validate the universality and cross-modal generalizability of this property across different modalities, model families, parameter scales, and tokenization schemes. Importantly, we treat changes in task performance as a behavioral readout of this property, acknowledging that redirected attention does not always result in improved task accuracy.

4.1 Experimental Setting

Our study utilizes a strictly controlled protocol where inputs are transformed *only* via letter case manipulation. Lexical content, word order, punctuation, and spacing are preserved identically to the original benchmark distributions to ensure that any observed shifts in model behavior are attributable solely to orthographic variation. For the cross-modal vision grounding experiments, we instantiate this protocol by embedding target object labels into generic template sentences to form an image description prompt. The label string within this prompt functions as the target span undergoing typographic alteration. We then isolate and quantify the internal cross-modal attention mass allocated by the model within the spatial region bounded by the ground-truth bounding box annotation corresponding to that label.

Models. We evaluate 13 models across five representative model families to ensure the property is not an artifact of a specific training regime or tokenizer:

- **LLaMA-3 Family** [17]: We use LLaMA-3.2-3B-Instruct [30] and LLaMA-3.1-8B-Instruct [29]. These models utilize the Byte-Pair Encoding (BPE)-based `tiktoken` tokenizer [33].
- **Gemma Family** [12,51,52]: We evaluate Gemma-3-1B-IT [14], Gemma-2-2B-IT [13], Gemma-3-4B-IT [15], and Gemma-4-E4B-IT [16]. These models employ a SentencePiece tokenizer with split digits, preserved whitespace, and byte-level encodings [24].
- **Mistral** [22]: We use Mistral-7B-Instruct-v0.3 [31], a widely adopted mid-scale model with a distinct SentencePiece [24] implementation that preserves case-related linguistic features.
- **Qwen Family** [38,40]: We use Qwen2.5-7B-Instruct [37], Qwen2.5-14B-Instruct [36], Qwen3-4B-Thinking-2507 [39], Qwen3-VL-4B-Instruct [42], and Qwen3-VL-2B-Thinking [41]. These models feature byte-level BPE (BBPE) tokenization [4] designed for multilingual robustness.
- **GPT**: We use the gpt-oss-20B model [34], a model specifically designed for advanced reasoning using the BPE-based `o200k_harmony` tokenizer.

This selection enables controlled comparison across tokenizer sensitivity to case (BPE vs. SentencePiece vs. BBPE), parameter scale (1B–20B), pretraining corpus composition, and between textual and visual modalities, critical for isolating whether case steering generalizes beyond a single model.

Benchmarks. We select three text understanding benchmarks and one vision-language benchmark that provide ground-truth answer spans or spatial bounding box annotations, allowing us to precisely target the typographic interventions. We evaluate factual and scientific reasoning using **MMLU-Pro** [56] and **ARC-Challenge** [8], reading comprehension with **SQuADv2** [43], and cross-modal visual grounding via **RefCOCOg** [27]. Multilingual and code generation evaluations on **XQuAD** [3] and **HumanEval** [7] are detailed in Appendix D.

4.2 Taxonomy of Typographic Interventions

To characterize the boundaries of LLM and VLM case sensitivity, we define a taxonomy of *static, deterministic interventions*. These interventions manipulate the contrast between the **target sequence** (the answer-relevant span or bounding box label) and the **context** (the remainder of the textual input). By varying the intensity and style of this contrast, we can determine whether the model responds to absolute casing (*e.g.*, uppercase) or relative salience (*e.g.*, case mismatch). All schemes produce strings $\tilde{X}$ with identical semantic content to the original input X, differing solely in surface-form typography.

Global Uniformity (U). These serve as baselines to measure the impact of removing all typographic cues.

- **U1: All-Caps**. Entire input in uppercase.
- **U2: All-Lowercase**. Entire input in lowercase.
- **U3: Title-Case**. Every word capitalized.

Target Emphasis (TE). These test the hypothesis that uppercase acts as an attentional attractor when placed against a de-emphasized background.

- **TE1 (Max-Contrast)**: Target in uppercase; context in lowercase.
- **TE2 (Pattern-Conflict)**: Target in uppercase; context in alternating case (*e.g.*, `wRoNg`).
- **TE3 (Natural-Context)**: Target in uppercase; context retains the original casing.

Target Title Case (TT). These probe whether subtler, grammatically conventional emphasis is sufficient to shift attention.

- **TT1 (Inverse-Contrast)**: Target in title case; context in uppercase.
- **TT2 (Standard-Emphasis)**: Target in title case; context in lowercase.
- **TT3 (Standard-Natural)**: Target in title case; context retains the original casing.

Target Alternating Case (TA). These test the model's response to highly unnatural, high-entropy typographic patterns.

- **TA1 (Alt-Upper)**: Target in alternating case (*e.g.*, `CoRrEcT`); context in uppercase.
- **TA2 (Alt-Lower)**: Target in alternating case; context in lowercase.
- **TA3 (Alt-Natural)**: Target in alternating case; context retains the original casing.

Adversarial De-Emphasis (ADE). Critical for causal validation, these interventions deliberately suppress the target span while emphasizing the context.

- **ADE1 (Target-Suppression)**: Target in lowercase; context in uppercase.
- **ADE2 (Target-Conflict)**: Target in lowercase; context in alternating case.
- **ADE3 (Target-Lower-Natural)**: Target in lowercase; context retains the original casing.

Latency and Tokenization Dynamics. The computational overhead of these surface-form variations is negligible (mean ≈ 0.0015 ms; see Tab. 17, App. E). While different casing can occasionally alter token counts due to tokenizer behavior, the consistency of results across diverse tokenizers confirms that the "casing effect" is a representational property rather than a byproduct of sequence length. By applying these interventions, we isolate letter case as a causal variable. This setup allows us to observe how the model's internal state reacts to surface-level changes without modifying the model's weights or lexical and visual input.

4.3 Characterizing Textual Attention Allocation Dynamics

Our empirical analysis begins by characterizing how typographic interventions modulate internal attention mass. We use Qwen2.5-7B-Instruct [37] as the discovery model. Results show that each intervention alters attention allocation, with magnitude varying substantially by pattern (*cf.* Figs. 2 and 9).

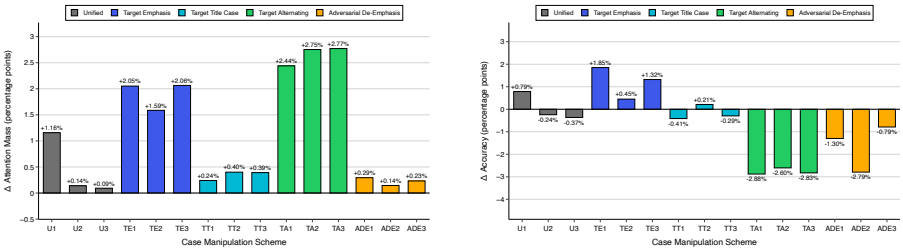
The most substantial shifts in attention toward the target span are elicited by **Target Alternating (TA)** schemes. Specifically, TA3 (alternating target, natural context) and TA2 (alternating target, lowercased context) yield the highest gains of +2.77 pp and +2.75 pp in mean attention mass, respectively. The negligible difference between TA2 and TA3 suggests that since natural English context is predominantly lowercase, the model responds primarily to the high-entropy alternating pattern of the target. Even when the context is uppercased (TA1), the target attracts a +2.44 pp gain, marking alternating case as the most potent attentional attribution shifter in our taxonomy.

Target Emphasis (TE) schemes, involving full uppercasing of the target, also yield notable shifts. TE3 (uppercase target, natural context) and TE1 (uppercase target, lowercase context) increase attention allocation by +2.06 pp and +2.05 pp, respectively. This effect is attenuated when the context is formatted in alternating case (TE2), dropping to +1.59 pp. This suggests that while uppercase is a strong attractor, its salience is partially diminished when the background context itself possesses high typographic entropy.

In contrast, unified patterns (U2, U3), title-case schemes (TT), and adversarial de-emphasis (ADE) interventions exert minimal influence, with attention shifts typically remaining below +0.4 pp. A notable exception is global uppercasing (U1), which induces a general gain of +1.16 pp, suggesting that the model treats all-caps text with a higher baseline of attentional intensity.

4.4 The Attention-Performance Divergence

A central finding of our study is the non-trivial relationship between attention allocation and downstream task performance. While alternating case (TA) is the most effective mechanism for concentrating attention, it serves as a *destructive attractor*: across all benchmarks, TA schemes consistently degrade downstream accuracy, with drops of up to −2.88 pp on the discovery model (*cf.* Fig. 11). This suggests that the high-entropy nature of alternating case, while salient to



(a) Relative Attention Mass on Target **(b)** Relative Task Accuracy

Fig. 2: Impact of typographic interventions on model behavior. Comparative analysis of relative attention mass allocated to the target span and downstream task accuracy across different case-steering schemes using Qwen2.5-7B-Instruct.

the attention mechanism, likely disrupts the model's ability to extract semantic meaning from the tokens (*cf*. App. B).

Conversely, **Target Emphasis (TE)** via uppercasing acts as a *productive attractor*. TE1 (uppercase target, lowercase context) achieves a performance gain of +1.85 pp. This indicates that the attention directed by uppercase lettering is utilized by the model, aligning with the hypothesis that LLMs internalize uppercase as a signal for importance rather than just a visual anomaly. Absolute values and detailed per-scheme breakdowns are provided in Table 9 in the Appendix.

4.5 Case Sensitivity in Vision-Language Models

Having established the casing effect on purely textual streams, we ask whether typographic salience crosses the modality boundary. We stress at the outset that we do *not* expect the multimodal setting to reproduce the text-only phenomenon verbatim: in text, casing concentrates attention *within* a single stream, and the resulting attention–performance divergence is an intra-modal effect. In VLMs, the same surface-form perturbation instead operates as a *cross-modal routing lever*. We therefore frame our multimodal results as a *partial* replication—the *family-level ordering* of attractors transfers across modalities, while its mechanism and downstream consequences differ in ways specific to multimodal processing. Concretely, a casing-induced steering effect is clearly present in VLMs, but it is less consistent than in the text-only regime: whereas the textual effect holds uniformly across every model and tokenizer family (*cf*. Sec. 4.6), the multimodal effect is robust only at the level of pattern-family ordering and the macroscopic modality axis, and becomes fragmented across individual models at the fine-grained spatial level.

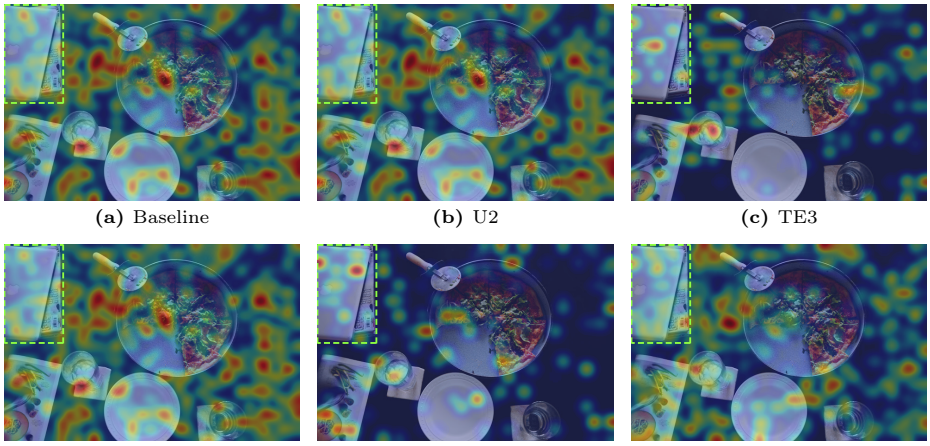
We evaluate our intervention taxonomy on the RefCOCOg visual grounding benchmark [27] across four VLMs spanning two families and both inference regimes: Qwen3-VL-4B-Instruct [42], Qwen3-VL-2B-Thinking [41], Gemma-3-4B-IT [15], and Gemma-4-E4B-IT (Thinking) [16]. Crucially, the intervention is applied *only* to the textual description prompt; the image is never altered.

Across the dataset, non-standard casing induces a coupled, dual-axis response (*cf*. Fig. 3): a *microscopic* concentration of the residual visual budget onto the target region (a mean +1.55 pp absolute shift, a relative +23.08% increase in target-region attention alignment), together with a *macroscopic* re-allocation away from the image as a whole (a mean −1.85 pp drop in whole-image attention mass, *i.e.*, a 12.41% relative disengagement from the visual stream in favour of the textual prompt). Consistent with reports of textual dominance in multimodal transformers [57], the dominant effect is macroscopic: casing primarily decides *how much* of the image is attended to, and secondarily *where*.

The family hierarchy transfers. Aggregating by pattern family reproduces the ordering observed on text. The Target Alternating (TA) family is again the strongest attractor, yielding the largest microscopic pull (+2.84 pp on the target box) and the largest macroscopic shift (+24.22% disengagement); Target Emphasis (TE) follows (+1.77 pp; +21.12%), while uniform (U), title-case (TT), and de-emphasis (ADE) families remain comparatively weak. This cross-modal persistence of the TA > TE > rest ordering indicates that the model treats high-entropy and uppercase prompt tokens as importance signals *regardless of the modality the answer lives in*. We caution, however, that the hierarchy is robust only at the family-aggregate level: as detailed in Appendix C.5 (Tab. 13), the per-model microscopic ordering is noisy—on the Qwen3-VL models, uniform uppercasing (U1), title-case contrast (TT1), and even de-emphasis (ADE1) rival TA1, whereas Gemma-3-4B-IT exhibits a sharper response (peaks of +6.23 pp under TA1 and +5.74 pp under TE2). The *macroscopic* axis is, by contrast, sign-consistent across all four models and constitutes the more reliable multimodal signature of the "casing effect" in VLMs.

Pattern-conflict drives modality disengagement. At the individual-pattern level, the pattern-conflicting configuration TE2 (uppercase target inside an alternating-case context) is the primary driver of modality disruption, producing the largest whole-image drain (−4.11 pp) and the strongest cross-model disengagement (a mean of 33.35%, and up to 43.74% on Qwen3-VL-2B-Thinking). When the surrounding prompt carries high orthographic entropy, processing concentrates on linguistic tokens at the expense of the visual feature maps (mechanistic detail in App. B). Uniform lowercasing (U2) minimises these shifts, perturbing spatial allocation by only +0.48 pp.

Reasoning reverses its role across modalities. The most salient multimodal-specific finding is a reversal of the "reasoning buffer" we identified on text. Whereas extended reasoning *attenuates* typographic sensitivity in text-only LLMs (*cf*. App. C.2), the visual reasoning phase *amplifies* macroscopic reliance on the text stream under typographic stress. Reasoning VLMs disengage from vision most strongly (Qwen3-VL-2B-Thinking: mean 14.88%, peak 43.74% under TE2; Gemma-4-E4B-IT: mean 12.92%, peak 28.83% under TA1), while the direct-inference models re-allocate more modestly (Gemma-3-4B-IT 11.20%, Qwen3-VL-4B-Instruct 10.64%). Conversely, direct-inference models are the more vul-



(a) Baseline **(b)** U2 **(c)** TE3

(d) TT3 **(e)** TA1 **(f)** ADE1

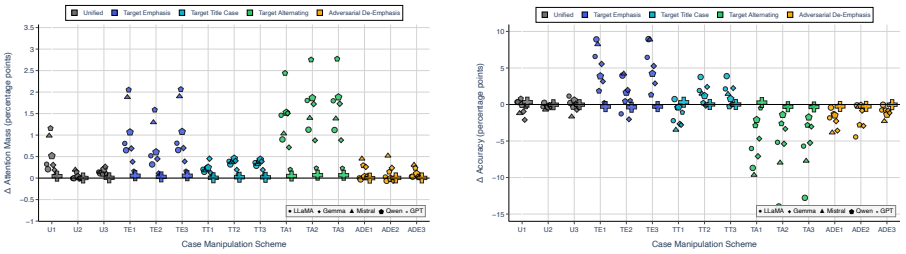
Fig. 3: Cross-modal attention steering via text-based typographic salience in VLMs shown with Grad-CAM visualizations of the Vision Transformer cross-attention maps on the RefCOCOg grounding benchmark with Qwen3-VL-2B-Thinking.

nerable to fine-grained *microscopic* steering. We interpret this as a division of labor: reasoning loops bias the macro-level modality choice toward textual coherence, while standard models remain exposed to raw spatial salience cues.

4.6 Cross-Model Generalization and Robustness

To verify that these behaviors are inherent properties of the transformer architecture rather than model-specific artifacts, we evaluate the intervention taxonomy across 13 models. Detailed quantitative results, including extended evaluations across diverse languages and specialized tasks, can be seen in Tables 4 to 12 as well as in Appendix D and F. A consistent picture emerges across both modalities, but with an important asymmetry in consistency: in the text-only setting the casing effect is *universal*, applying across all models and tokenizer families, whereas in VLMs the same effect arises only *partially*—robust at the family-aggregate and macroscopic level, yet fragmented across individual models.

– **Textual Observation (Universal Effect)**: As shown in Figure 4, non-reasoning LLMs exhibit consistent patterns of attention shift and corresponding task-accuracy variations across all schemes, albeit with differing magnitudes. This consistency is *universal* in non-reasoning models: it holds across every evaluated model and across all three tokenizer families (BPE, SentencePiece, and BBPE; *cf*. App. F), confirming that the effect is a representational property rather than a tokenizer artifact. In contrast, reasoning models deviate from this trend, showing near-zero sensitivity in both attention allocation and downstream task accuracy (*cf*. App. C.2).



(a) Relative Attention Mass on Target **(b)** Relative Task Accuracy

Fig. 4: Generalization of typographic interventions across different models. Comparative analysis of relative attention mass allocated to the target span and downstream task accuracy across different case-steering schemes and model architectures. Symbols encode the primary model family, while their size encodes model size.

- **Reasoning Influence**: The deliberative reasoning process in text-only LLMs largely mitigates shifts in attention mass and exhibits minimal variation in downstream task accuracy (typically below ±0.5 pp). This suggests that the "thinking" phase acts as a semantic realignment layer that filters out typographic salience in favor of logical coherence (*cf.* App. C.2).
- **Consistent Failure Modes**: In non-reasoning models, high-entropy casing acts as a destructive attractor; alternating case (TA) consistently degrades performance, reaching a peak loss of −13.96 pp in LLaMA-3.1-8B-Instruct. Adversarial de-emphasis (ADE) schemes similarly confirm that misdirecting attention through casing systematically lowers task accuracy.
- **Utility of Uppercasing**: Across all standard architectures, TE1 and TE3 remain the only interventions that reliably yield performance gains. These schemes achieve mean accuracy increases of up to +8.95 pp, suggesting that standard LLMs have internalized capitalization as valid importance signal.
- **Partial Transfer to VLMs**: The textual hierarchy reproduces *partially* in the visual domain. Text-level casing acts as a zero-shot cross-modal lever, increasing the attention mass that falls inside the target bounding box by a relative +23.08% (a mean +1.55 pp absolute shift), with the family ordering (TA > TE > rest) preserved in aggregate. Unlike the universal textual effect, however, the per-model spatial response is fragmented: direct-inference pipelines are the most locally sensitive (peaking at a +6.23 pp spatial attention surge in Gemma-3-4B-IT), whereas the Qwen3-VL models exhibit no clear family ordering at the fine-grained level (*cf.* Tabs. 13 and 14).
- **Macroscopic Modality Shifts**: The axis that is sign-consistent across all four VLMs is macroscopic. Non-standard casing redirects attention away from the image toward the text prompt, yielding a global +12.41% disengagement from the visual stream (a −1.85 pp whole-image drop). This shift is amplified by visual reasoning—reversing the text-only "reasoning buffer"—and peaks at a 43.74% total visual attention drop in Qwen3-VL-2B-Thinking.

5 Conclusion

In this paper, we presented a systematic empirical study characterizing letter case as a fundamental modulator of internal attention in both LLMs and VLMs. By isolating typographic variation from semantic and lexical content, we established that casing is not merely a tokenization artifact but a latent property inherent to the transformer's pretraining. Our results across seven non-reasoning LLMs demonstrate that interventions such as formatting target spans in uppercase reliably concentrate attention mass and improve downstream performance by up to +8.95 pp without model modification, an effect that holds across every evaluated non-reasoning model and tokenizer family. Extending this paradigm to the visual domain, we find that text-level casing transfers across the modality boundary partially: the family-level ordering of attractors is preserved in aggregate, but casing reorganizes cross-modal attention along two coupled axes—shifting overall allocation away from the image (the dominant effect), and secondarily concentrating the residual visual budget on the target region.

A central contribution is characterizing the non-trivial relationship between attention allocation and task utility. We identified a divergence between *productive* attractors (uppercase), which enhance downstream performance, and *destructive* attractors (alternating case), which elicit the strongest attention shifts but significantly degrade performance. Crucially, our analysis uncovers a boundary condition that *reverses* across modalities: whereas the text-only "thinking" phase acts as a semantic buffer that filters out typographic salience, the visual reasoning phase instead amplifies macro-level reliance on the textual stream. Under typographic stress, reasoning VLMs exhibit a pronounced shift away from visual features, whereas direct-inference models remain vulnerable to localized spatial steering.

Limitations and Future Work. Our evaluation focused on languages where letter casing is defined; exploring this property in non-Latin scripts and applying causal mediation analysis would provide deeper mechanistic insights. Additionally, developing automated pipelines that leverage this property in textual and visual contexts represents a promising path toward efficient, black-box attention steering. Ultimately, this work establishes typographic salience as a training-free attention-steering lever across text and vision domains.

Acknowledgments

This research was funded, in part, by the U.S. Government under ARPA-H contract 1AY2AX000062 and DARPA contract HR0011-24-9-0429. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

1. Alpay, F., Alpay, T.: Xml prompting as grammar-constrained interaction: Fixed-point semantics, convergence guarantees, and human-ai protocols (2025), https://arxiv.org/abs/2509.08182
2. Anonymous: Spectral attention steering for prompt highlighting. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=XfLvGIFmAN
3. Artetxe, M., Ruder, S., Yogatama, D.: On the cross-lingual transferability of monolingual representations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. p. 4623–4637. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.421, http://dx.doi.org/10.18653/v1/2020.acl-main.421
4. Bai, J., Bai, S., Chu, Y., et al.: Qwen technical report (2023), https://arxiv.org/abs/2309.16609
5. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
6. Bu, R., Zhong, H., Chen, W., Li, Y.: Value-state gated attention for mitigating extreme-token phenomena in transformers (2026), https://arxiv.org/abs/2510.09017
7. Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H.P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F.P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W.H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A.N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., Zaremba, W.: Evaluating large language models trained on code (2021), https://arxiv.org/abs/2107.03374
8. Chollet, F.: On the measure of intelligence (2019), https://arxiv.org/abs/1911.01547
9. Cutter, M.G., Martin, A.E., Sturt, P.: Capitalization interacts with syntactic complexity. Journal of Experimental Psychology: Learning, Memory, and Cognition **46**(6), 1146 (2020)
10. Davarmanesh, P., Wilson, A., Radhakrishnan, A.: Efficient and accurate steering of large language models through attention-guided feature learning (2026), https://arxiv.org/abs/2602.00333
11. Fournet, C., Mirault, J., Perea, M., Grainger, J.: Effects of letter case on processing sequences of written words. Journal of experimental psychology: learning, memory, and cognition **48**(12), 1995 (2022)
12. Google Developers: Gemma core models. https://ai.google.dev/gemma/docs/core (2026), accessed: 2026-06-21

13. Google Team: Gemma-2-2b-it. https://huggingface.co/google/gemma-2-2b-it (2024)
14. Google Team: Gemma-3-1b-it. https://huggingface.co/google/gemma-3-1b-it (2025)
15. Google Team: Gemma-3-4b-it. https://huggingface.co/google/gemma-3-4b-it (2025)
16. Google Team: Gemma-4-e4b-it. https://huggingface.co/google/gemma-4-E4B-it (2026)
17. Grattafiori, A., Dubey, A., Jauhri, A., et al.: The llama 3 herd of models (2024), https://arxiv.org/abs/2407.21783
18. Guardieiro, V., Stein, A., Khare, A., Wong, E.: Instruction following by boosting attention of large language models (2025), https://arxiv.org/abs/2506.13734
19. Hernandez, E., Li, B.Z., Andreas, J.: Inspecting and editing knowledge representations in language models (2024), https://arxiv.org/abs/2304.00740
20. Hsieh, C.Y., Chuang, Y.S., Li, C.L., Wang, Z., Le, L., Kumar, A., Glass, J., Ratner, A., Lee, C.Y., Krishna, R., Pfister, T.: Found in the middle: Calibrating positional attention bias improves long context utilization. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 14982–14995. Association for Computational Linguistics, Bangkok, Thailand (08 2024). https://doi.org/10.18653/v1/2024.findings-acl.890
21. Inhoff, A.W., Starr, M., Shindler, K.L.: Is the processing of words during eye fixations in reading strictly serial? Perception & Psychophysics **62**(7), 1474–1484 (2000). https://doi.org/10.3758/BF03212147, https://doi.org/10.3758/BF03212147
22. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023), https://arxiv.org/abs/2310.06825
23. Klinke, T., Christ, M., Fadl, N., Lamerz, C., Langner, T.: The effects of letter capitalization in advertising headlines. Journal of Marketing Communications **0**(0), 1–23 (2024). https://doi.org/10.1080/13527266.2024.2401393
24. Kudo, T., Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing (2018), https://arxiv.org/abs/1808.06226
25. Lee, B.W., Padhi, I., Ramamurthy, K.N., Miehling, E., Dognin, P., Nagireddy, M., Dhurandhar, A.: Programming refusal with conditional activation steering (2025), https://arxiv.org/abs/2409.05907
26. Li, Z., Geng, B., Xiong, J., He, Y., Hu, Y., Chen, J., Chen, D., Chang, X., Zhang, L., Mo, L., Li, C., Yuan, C., Sun, Z.: Ctr-sink: Attention sink for language models in click-through rate prediction (2025), https://arxiv.org/abs/2508.03668
27. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A., Murphy, K.: Generation and comprehension of unambiguous object descriptions (2016), https://arxiv.org/abs/1511.02283
28. Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and editing factual associations in gpt. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 17359–17372. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf
29. Meta: Llama-3.1-8b-instruct. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct (2024)

30. Meta: Llama-3.2-3b-instruct. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct (2024)
31. Mistral AI: Mistral-7b-instruct-v0.3. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3 (2024)
32. Mitchell, E., Lin, C., Bosselut, A., Finn, C., Manning, C.D.: Fast model editing at scale (2022), https://arxiv.org/abs/2110.11309
33. OpenAI: tiktoken: A fast bpe tokeniser for use with openai's models. https://github.com/openai/tiktoken (2023)
34. OpenAI: gpt-oss-120b gpt-oss-20b model card (2025), https://arxiv.org/abs/2508.10925
35. Phan, B., Havasi, M., Muckley, M., Ullrich, K.: Understanding and mitigating tokenization bias in language models (2024), https://arxiv.org/abs/2406.16829
36. Qwen Team: Qwen2.5-14b-instruct. https://huggingface.co/Qwen/Qwen2.5-14B-Instruct (2024)
37. Qwen Team: Qwen2.5-7b-instruct. https://huggingface.co/Qwen/Qwen2.5-7B-Instruct (2024)
38. Qwen Team: Qwen2.5 technical report (2025), https://arxiv.org/abs/2412.15115
39. Qwen Team: Qwen3-4b-thinking-2507. https://huggingface.co/Qwen/Qwen3-4B-Thinking-2507 (2025)
40. Qwen Team: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388
41. Qwen Team: Qwen3-vl-2b-thinking. https://huggingface.co/Qwen/Qwen3-VL-2B-Thinking (2025)
42. Qwen Team: Qwen3-vl-4b-instruct. https://huggingface.co/Qwen/Qwen3-VL-4B-Instruct (2025)
43. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for squad (2018), https://arxiv.org/abs/1806.03822
44. von Rütte, D., Anagnostidis, S., Bachmann, G., Hofmann, T.: A language model's guide through latent space (2024), https://arxiv.org/abs/2402.14433
45. Sinclair, A., Jumelet, J., Zuidema, W., Fernández, R.: Structural persistence in language models: Priming as a window into abstract language representations. Transactions of the Association for Computational Linguistics **10**, 1031–1050 (09 2022). https://doi.org/10.1162/tacl_a_00504, https://doi.org/10.1162/tacl_a_00504
46. Singh, A., Fry, A., Perelman, A., et al.: Openai gpt-5 system card (2025), https://arxiv.org/abs/2601.03267
47. Sinii, V., Gorbatovski, A., Cherepanov, A., Shaposhnikov, B., Balagansky, N., Gavrilov, D.: Steering llm reasoning through bias-only adaptation (2025), https://arxiv.org/abs/2505.18706
48. Slattery, T.J., Schotter, E.R., Berry, R.W., Rayner, K.: Parafoveal and foveal processing of abbreviations during eye fixations in reading: Making a case for case. Journal of Experimental Psychology: Learning, Memory, and Cognition **37**(4), 1022 (2011)
49. Stolfo, A., Balachandran, V., Yousefi, S., Horvitz, E., Nushi, B.: Improving instruction-following in language models through activation steering (2025), https://arxiv.org/abs/2410.12877
50. Tamayo, D., Gonzalez-Agirre, A., Hernando, J., Villegas, M.: Mass-editing memory with attention in transformers: A cross-lingual exploration of knowledge. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 5831–5847. Association for Computational

Linguistics, Bangkok, Thailand (8 2024). https://doi.org/10.18653/v1/2024.findings-acl.347, https://aclanthology.org/2024.findings-acl.347/
51. Team, G., Kamath, A., Ferret, J., et al.: Gemma 3 technical report (2025), https://arxiv.org/abs/2503.19786
52. Team, G., Riviere, M., Pathak, S., et al.: Gemma 2: Improving open language models at a practical size (2024), https://arxiv.org/abs/2408.00118
53. Turner, A.M., Thiergart, L., Leech, G., Udell, D., Vazquez, J.J., Mini, U., MacDiarmid, M.: Steering language models with activation engineering (2024), https://arxiv.org/abs/2308.10248
54. Venkateswaran, P., Contractor, D.: Spotlight your instructions: Instruction-following with dynamic attention steering (2026), https://arxiv.org/abs/2505.12025
55. Vergara-Martínez, M., Perea, M., Leone-Fernandez, B.: The time course of the lowercase advantage in visual word recognition: An erp investigation. Neuropsychologia **146**, 107556 (2020). https://doi.org/https://doi.org/10.1016/j.neuropsychologia.2020.107556
56. Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., Chen, W.: Mmlu-pro: A more robust and challenging multi-task language understanding benchmark (2024), https://arxiv.org/abs/2406.01574
57. Wu, H., Tang, M., Zheng, X., Jiang, H.: When language overrules: Revealing text dominance in multimodal large language models (2025), https://arxiv.org/abs/2508.10552
58. Wu, X., Wang, Y., Jegelka, S., Jadbabaie, A.: On the emergence of position bias in transformers (2025), https://arxiv.org/abs/2502.01951
59. Zhang, Q., Singh, C., Liu, L., Liu, X., Yu, B., Gao, J., Zhao, T.: Tell your model where to attend: Post-hoc attention steering for llms (2024), https://arxiv.org/abs/2311.02262
60. Zhang, Q., Yu, X., Singh, C., Liu, X., Liu, L., Gao, J., Zhao, T., Roth, D., Cheng, H.: Model tells itself where to attend: Faithfulness meets automatic attention steering (2024), https://arxiv.org/abs/2409.10790